14.3 An 8MB Level-3 Cache in 32nm SOI with Column-Select Aliasing

Don Weiss[1], Michael Dreesen[1], Michael Ciraula[1], Carson Henrion[1], Chris Helt[1], Ryan Freese[1], Tommy Miles[1], Anita Karegar[1], Russell Schreiber[2], Bryan Schneller[1], John Wuu[1]

[1]AMD, Fort Collins, CO,
[2]AMD, Austin, TX

High-performance multi-core processors require efficient multi-level cache hierarchies to meet high-bandwidth data requirements. Because level-3 (L3) cache is typically the largest cache on the die, the drive to lower cost places pressure on density, yields, and test time. Performance-per-watt goals and total power constraints also compel a variety of circuit techniques to reduce power. The next-generation server processor codenamed "Orochi", implemented on a 32nm high-k metal-gate SOI process with 11 metal layers, consists of four 2-core modules using AMD's next-generation architecture, code named "Bulldozer", with 2MB of dedicated L2 cache per module and an 8MB shared L3 cache [1].

The L3 cache is divided into 4 independent 2MB subcaches, shown in Fig. 14.3.1, with each subcache further divided into 4 banks. The interface to the subcache consists of the following ports: 2 tag read, 2 tag write, 2 data read and 1 data write. Concurrent operations on different ports are supported when accesses do not target the same bank. The data portion of the subcache is divided into 4 sequential regions, each running one phase behind the previous region and containing 1/4 of the 512b cache line. The combination of half-cycle delay for control signals reaching a region and half-cycle delay for read data crossing that region means a cache line is returned in a burst of 4 cycles during a read. Similarly, the subcache requires 4 cycles to write a full cache line. A sequential region contains 8 64KB macros, each containing 64 data, 6 ECC and 2 repair I/O's, with 2 macros in each sequential region accessed per operation. The macro operates in a flow-through manner, illustrated in Fig. 14.3.2, to enable high speed and area efficiency while reducing clock power [2,3].

The data array uses short bitlines containing 32 cells for robust read stability. To achieve high density, a compact single-ended read circuit is used for every 16 columns, with 8 columns on each of its 2 sides (Fig. 14.3.3). Pre-MUX read sensing achieves high-speed read operations and low-voltage performance. Writes are performed with true and complement data driven through an 8:1 NFET write-MUX with cross-coupled PFETs on the bitlines. Traditionally, separate read and write 8:1 MUXes would require 16 column-select signal tracks in layout, consuming additional area. Using the same column-select signals to drive the read and write MUXes could halve the number of required tracks, but would defeat the purpose of pre-MUX sensing because the load inside the write-MUX would be exposed to the read bitline, dramatically reducing speed.

A solution dubbed column-select aliasing (CSA) uses only 8 wires to drive both MUXes by aliasing odd and even pairs of column-selects for reads and writes. Figure 14.3.4 illustrates the concept with a pair of columns as an example. CSEL<0> is used as "column 0 select" for reads and "column 1 select" for writes; CSEL<1> is used as "column 1 select" for reads and "column 0 select" for writes. Therefore, the bitline selected for a read is isolated from the load inside the write-MUX to maintain high speed, while the same set of column select signals can be reused for write operations. The local write drivers are tri-stated during read operations to avoid contention with the SRAM cells on the aliased columns. Because of its significant area savings, CSA was also used in the L3 tag and least-recently used (LRU) arrays. Different column select enhancements were applied to other areas of the chip. For example, the L2 tag employs a shared, oversized write driver to simultaneously write all columns of a 4:1 MUX configuration in an invalidate operation which shortens the invalidate time.

Power reduction is another L3 cache design priority, and this design employs several techniques to reduce static power. For example, all of the large wordline and column select drivers in data macros are power-gated in groups of eight, accounting for 28% data macro static power reduction, or 20% reduction in overall subcache static power. Similarly, tag wordline drivers are power-gated to reduce the overall subcache static power by another 6% (Fig. 14.3.5). In traditional SRAM designs, bitlines remain precharged at V_{DD} when bitcells are not accessed, resulting in subthreshold leakage current into the logical 0 sides of the SRAM cells and junction leakage current from the N+ regions connected to the bitlines. By allowing the bitline voltages to float, these leakage components are decreased, reducing overall standby power. 18% power savings have been reported in 32nm bulk SRAM using floating-bitline mode (FBM) [4], while this design's 10% standby power reduction due to FBM represents a contrasting data point for 32nm SOI SRAM that has negligible junction leakage. V_{DD} on the read NAND gates is gated in FBM, which protects the NAND gate from crowbar current while the inputs are floating. Headers are not used for the cross-coupled bitline PFETs to keep area small, so as a compromise only half the bitlines receive the full benefits of FBM. In addition to power gating, high-V_T transistors are used throughout the design to reduce power further, and the SRAM bitcell is chosen for its use of low-leakage SRAM transistors at the expense of larger bitcell area. The combination of these techniques reduces the leakage power enough to eliminate the need for more complex and costly solutions, such as array sleep or regulated supplies.

Along with density and power, focus is also placed on yield and testability enhancements. Traditional row and column redundancies in each macro usually have several drawbacks, including requiring multiple BIST passes for identifying and programming replacement elements, inefficient use of area by including redundancy storage and control circuitry in all macros, and increased pressure on timing paths by placing the redundancy comparison and data insertion circuitry in critical paths. To address these drawbacks, this design uses centralized row and column redundancy blocks for its data macros. Redundant row data is stored in independent latch arrays separate from the data macros, and each row redundancy macro can repair any 12 rows of data across 16 data macros. The redundant column data is stored in each data macro; however, the repair address and CAM are stored in a separate column redundancy macro. Each column redundancy macro can repair a total of 24 columns across 16 data macros, replacing up to two columns per data macro. All redundancy MUXes are centrally located at the subcache interface. This centralized redundancy scheme offers lower area overhead, removes the speed penalty typically associated with row redundancy, and provides more flexibility in utilizing available repair elements. In addition, the centralized redundancy scheme reduces test time by requiring only 1 BIST pass for the data region, rather than needing additional passes after repairs are made.

This L3 cache is designed to operate above 2.3GHz at 1.1V, and its key technology, density, and bandwidth features are summarized in Fig. 14.3.6.

Acknowledgements:
The authors thank Mike Leary, Bill Hughes, Gregg Donley, Benjamin Tsien, Steve Foster, Sho-Chien Kang, Pouya Razavi, Nick Fournier, John Chan, the L2 team, the layout team, and GLOBALFOUNDRIES for their contributions and support.

References:
[1] T. Fischer, et al., "Design Solutions for a 32nm SOI 2-core processor module in an 8-core CPU," *ISSCC Dig. Tech. Papers*, Feb., 2011
[2] J. Dorsey, "An Integrated Quad-Core Opteron Processor," *ISSCC Dig. Tech. Papers*, pp. 102-103, Feb., 2007
[3] J. Wuu, et al., "The Asynchronous 24MB On-Chip Level-3 Cache for a Dual-Core Itanium-Family Processor," *ISSCC Dig. Tech. Papers*, pp. 488-489, Feb., 2005
[4] Y. Wang, et al., "A 4.0 GHz 291Mb Voltage-Scalable SRAM Design in 32nm High-K Metal-Gate CMOS with Integrated Power Management," *ISSCC Dig. Tech. Papers*, pp. 456-457, Feb., 2009.

 978-1-61284-302-5/11/$26.00

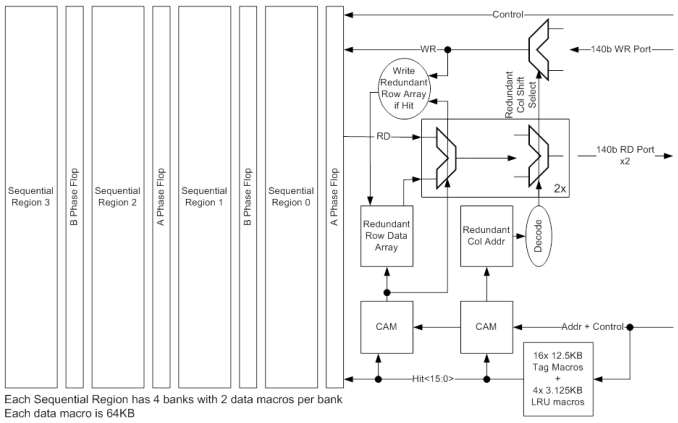

Figure 14.3.1: 2MB Subcache Organization.

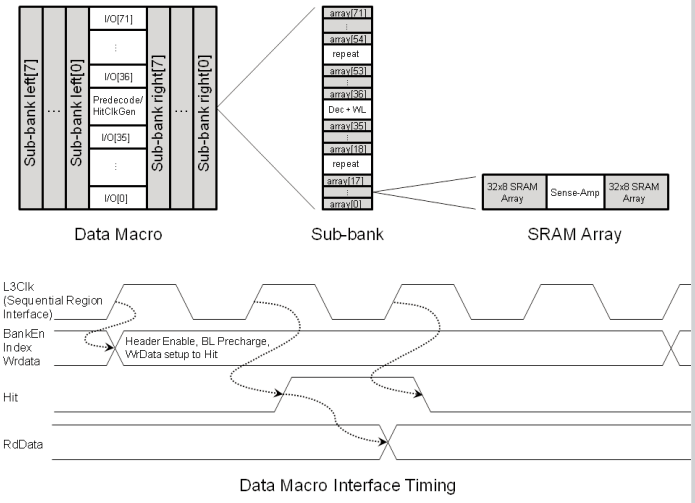

Figure 14.3.2: Data Macro Organization and Timing.

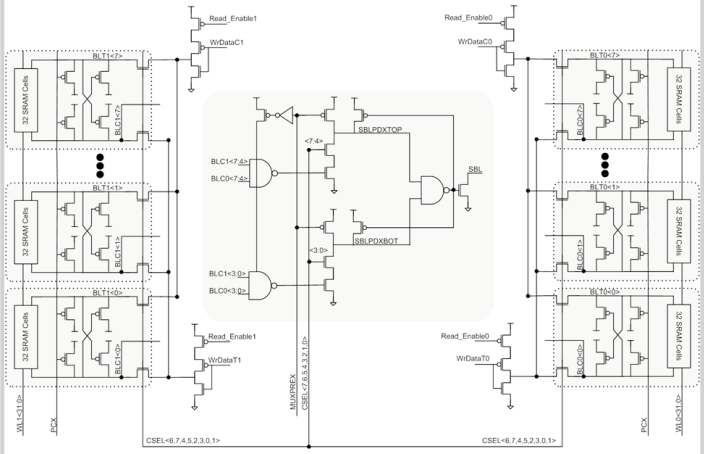

Figure 14.3.3: Sensing Circuit.

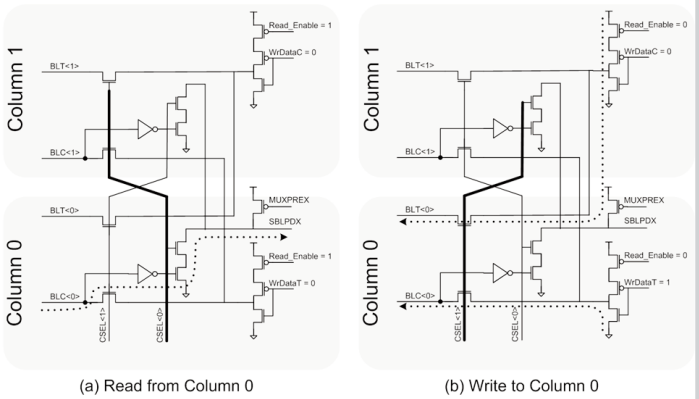

Figure 14.3.4: Column Select Aliasing Illustration.

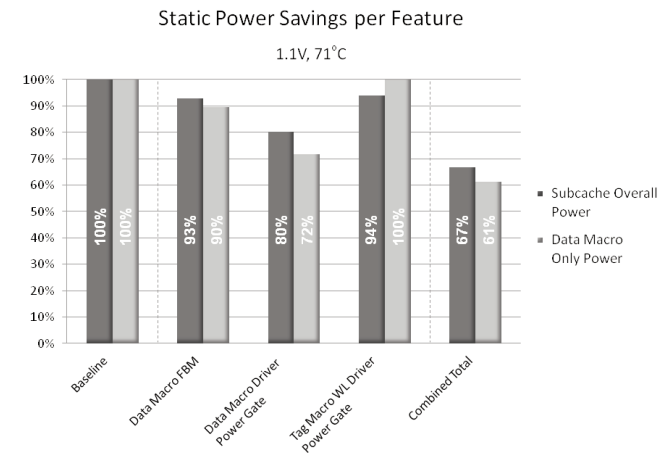

Figure 14.3.5: Static Power Reduction Features Summary.

Technology	32nm PD-SOI
Technology Features	High-K metal-gate, strained silicon
Metallization	Cu, 11 layers
Cache Size	8 MB
Set Associativity	16, 32, or 64 way
Line Size	512 bits
Data Read Bandwidth (2.4GHz peak)	307 GB/s
Data Write Bandwidth (2.4GHz peak)	154 GB/s
Memory Cell Area	0.258 um^2
Data Macro Density	3.6 mm^2/MB
Subcache Density including Tag, LRU, Redundancy, BIST, and Control	4.8 mm^2/MB

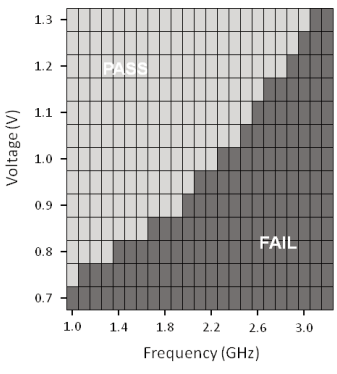

Figure 14.3.6: Feature Summary and Shmoo Plot.

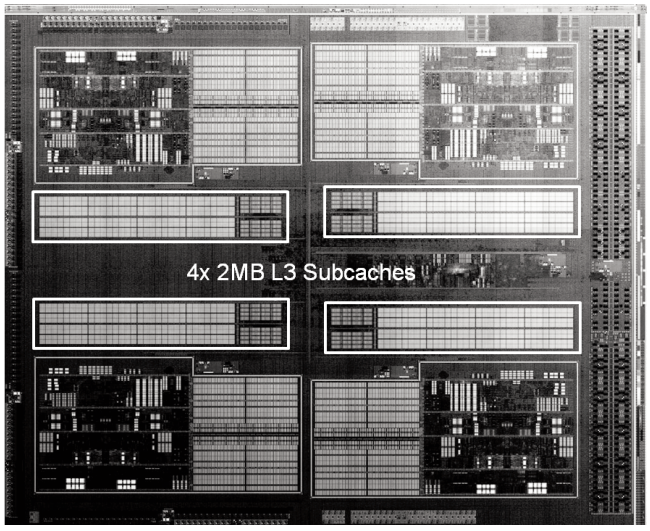

Figure 14.3.7: Die Photo.

978-1-61284-302-5/11/$26.00